Statistical Algorithms and a Lower Bound for Planted Clique

Vitaly Feldman Elena Grigorescu*† Lev Reyzin‡†

Santosh S. Vempala† Ying Xiao†

vitaly@post.harvard.edu
IBM Almaden Research Center
San Jose, CA 95120

{elena,lreyzin,vempala,yxiao32}@cc.gatech.edu
School of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

We develop a framework for proving lower bounds on computational problems over distributions, including optimization and unsupervised learning. Our framework is based on defining a restricted class of algorithms, called *statistical algorithms*, that instead of accessing samples from the input distribution can only obtain an estimate of the expectation of any given function on a sample drawn randomly from the input distribution. Our definition captures many natural algorithms used in theory and practice, e.g. moments-based methods, local search, MCMC and simulated annealing. Our techniques are inspired by (and generalize) the statistical query model in learning theory, which captures the complexity of PAC learning using essentially all known learning methods [Kearns, 1998].

For specific well-known problems over distributions, we give lower bounds on the complexity of *any* statistical algorithm. These include an exponential lower bounds for moment maximization in $\mathcal{R}^n$, and a nearly optimal lower bound for detecting planted clique distributions when the planted clique has size $O(n^{1/2-\delta})$ for any constant $\delta > 0$. Variants of the latter problem have been assumed to be hard to prove hardness for other problems and for cryptographic applications. Our lower bounds provide concrete evidence supporting these assumptions.

*This material is based upon work supported by the National Science Foundation under Grant #1019343 to the Computing Research Association for the CIFellows Project.

†Research supported in part by NSF awards AF-0915903 and AF-0910584.

‡Research supported by a Simons Postdoctoral Fellowship.

1 Introduction

Our primary motivation is to establish computational lower bounds on a set of well-known search and optimization problems defined over distributions that can be sampled. The traditional approach to this is based on reductions to problems conjectured to be intractable. Here we present a new approach: we show that a broad class of algorithms, which we refer to as *statistical algorithms*, must have high asymptotic complexity, unconditionally.

Our definition encompasses many well-known algorithms, such as EM [Dempster et al., 1977], local search, MCMC optimization [Tanner and Wong, 1987, Gelfand and Smith, 1990], simulated annealing [Kirkpatrick et al., 1983, Černý, 1985], as well as first and second order methods for linear/convex optimization, e.g. Dunagan and Vempala [2008]. We define this class of algorithms and show they must have high complexity for problems such as detecting large planted cliques or planted dense subgraphs, maximizing a polynomial over the unit sphere, maximum satisfiability, etc. These results rule out many natural approaches to solving these problems in theory and provide some practical guidance about when not to use popular and generic heuristics such as EM or simulated annealing. Our work also serves to highlight the question: what nonstatistical algorithms exist for search and optimization problems?

The inspiration for our model comes from the *statistical query (SQ)* model in learning theory [Kearns, 1998], where any algorithm that is based only on statistical queries must have complexity that grows with the statistical query dimension of the hypothesis class being learned [Blum et al., 1994]. In particular, this rules out polynomial-time SQ algorithms for learning parities from the uniform distribution on $\{-1, 1\}^n$. Our definition generalizes SQ algorithms which are known to capture almost all efficient algorithms for learning. Before we define our model precisely, we mention two specific motivating problems.

Detecting Planted Cliques. In the standard planted clique problem, we are given a graph G whose edges are generated by starting with a random graph $G_{n,1/2}$, then "planting" (adding edges to make) a clique on k vertices. Jerrum [1992] introduced the planted clique problem as a potentially easier variant of the classical problem of finding the largest clique in a random graph. A random graph $G_{n,1/2}$ contains a clique of size $2 \log n$ with high probability, and a simple greedy algorithm can find one of size $\log n$, and it appears hard to find one of size $(1 + \epsilon) \log n$ for any $\epsilon > 0$. Planting a larger clique should make it easier to find one. The smallest k for which such a clique can be detected in polynomial time is $\Omega(\sqrt{n})$ [Alon et al., 1998, McSherry, 2001], using an eigenvector-based algorithm. For $k \geq c\sqrt{n \log n}$, simply picking vertices of large degrees suffices [Kucera, 1995]. One intriguing aspect of this problem is that for any k, there is a quasipolynomial algorithm: guess $2 \log n$ vertices from the clique and take all their common neighbors.

Some evidence toward the hardness of the problem was shown by Jerrum [1992] who proved that a specific approach using a Markov chain cannot be efficient for small k. The problem has been used to generate cryptographic primitives [Juels and Peinado, 2000], as well as demonstrate the hardness of finding approximate Nash equilibria of certain games [Hazan and Krauthgamer, 2011, Minder and Vilenchik, 2009]. Bipartite versions of the planted clique problem have also been extensively studied. Here a bipartite clique is planted in a random bipartite graph. A version of the bipartite planted clique problem has been used as a hard problem for cryptographic applications [Applebaum et al., 2010]. We now define the planted bipartite clique problem formally.

Problem 1 (planted bipartite k-clique)**.** *For $1 \leq k \leq n$, let $S \subseteq \{1, 2, \ldots, n\}$ be a set of k vertex indices and D_S be a distribution over $\{0, 1\}^n$ such that when $x \sim D_S$, with probability $1 - (k/n)$*

*the entries of x are chosen uniformly and independently from $\{0,1\}$, and with probability k/n the k coordinates in S are set to 1 and the rest are chosen uniformly and independently from $\{0,1\}$. The **planted bipartite k-clique** problem is to find the unknown subset S given access to samples from D_S.*

One can view the vectors x as adjacency vectors of a random bipartite graph with n vertices on one side and a planted bipartite clique with an expected k/n fraction of vertices on either side. This formulation captures the traditional bipartite planted clique problem when exactly n examples are drawn from D. In addition to planted clique, our lower bounds will also apply to planted dense subgraphs — here the probability of a coordinate in S being 1 is $q > 1/2$. Known algorithms for these problems require cliques (or dense subgraphs) of size $k = \Omega(\sqrt{n})$. Our main result for this problem is a nearly matching lower bound for any statistical algorithm.

Moment Maximization. Our second example is an optimization problem defined as follows.

Problem 2 (moment maximization)**.** *Let D be a distribution over $[-1,1]^n$ and let $r \in Z_+$. The **moment maximization** problem is to find a unit vector u^* that maximizes the expected r'th moment of the projection of D to u^*, i.e.,*

$$u^* = \arg \max_{u \in \mathcal{R}^n : \|u\|=1} \mathop{\mathbf{E}}_{x \sim D} [(u \cdot x)^r].$$

The complexity of finding approximate optima is interesting as well. For $r = 2$, an optimal vector simply corresponds to the principal component of the distribution D and can be found by the singular value decomposition. For higher r, there are no efficient algorithms known, and the problem is NP-hard for some distributions [Brubaker, 2009, Hillar and Lim, 2009]. It can be viewed as finding the 2-norm of an r'th order tensor (the moment tensor of D). For $r = 3$, Frieze and Kannan [2008] give a reduction from finding a planted clique in a random graph to this tensor norm maximization problem; this was extended to general r in Brubaker and Vempala [2009]. Specifically, they show that maximizing the r'th moment (or the 2-norm of an r'th order tensor) allows one to recover planted cliques of size $\tilde{\Omega}(n^{1/r})$.

For moment maximization over a distribution that can be sampled, it is natural to consider the following type of optimization algorithm: start with some unit vector u, then estimate the gradient at u (via samples), and move along that direction staying on the sphere; repeat to reach a local maximum. Unfortunately, over the unit sphere, the expected r'th moment function can have (exponentially) many local maxima even for simple distributions. A more sophisticated approach [Kannan] for both problems is through Markov chains or simulated annealing; it attempts to sample unit vectors from a distribution on the sphere which is heavier on vectors that induce a higher moment, e.g., u is sampled with density proportional to $e^{f(u)}$ where $f(u)$ is the expected r'th moment along u. This could be implemented by a Markov chain with a Metropolis filter [Metropolis et al., 1953, Hastings, 1970] ensuring a proportional steady state distribution. If the Markov chain were to mix rapidly, that would give an efficient approximation algorithm because sampling from the steady state likely gives a vector of high moment. At each step, all one needs is to be able to estimate $f(u)$, which can be done by sampling from the input distribution.

As we will see presently, these approaches fall under a class of algorithms we call *statistical algorithms*, and they will all have provably high complexity and nearly matching upper bounds.

2 Definitions and overview

We now describe our model, approach for proving lower bounds and some applications in detail.

2.1 Model

The statistical query learning model of Kearns [1998] is a restriction of the PAC model [Valiant, 1984]. It captures algorithms that rely on empirical estimates of statistical properties of random examples of an unknown function instead of individual random examples (as in the PAC model of learning). Here a statistical property refers to the expectation of any boolean function of an example with respect to the unknown distribution of examples.

In the same spirit, for general search, decision and optimization problems over a distribution, we define statistical algorithms as algorithms that do not see samples from the distribution but instead have access to estimates of the expectation of any bounded function of a sample from the distribution.

Definition 1 (statistical algorithms)**.** *Let D be the input distribution over the domain X. We say that an algorithm is **statistical** if it does not have direct access to samples from D, but instead makes calls to an oracle* STAT_D*, which takes as input any function $h \in \mathcal{H} : X \to [-1, 1]$ and a* tolerance *parameter $\tau > 0$.* $\mathrm{STAT}_D(h, \tau)$ *returns a value*

$$v \in [h(D) - \tau, h(D) + \tau].$$

The most natural realization of a STAT_D oracle is one that computes h on $O(1/\tau^2)$ random samples from D and returns their average. In fact, as we will show later, $1/\tau^2$ roughly corresponds to the sample complexity of a (usual) algorithm whereas the number of queries roughly corresponds to the running time complexity.

The general algorithmic techniques mentioned earlier can all be expressed in this model in a relatively straightforward way. We would also like to note that in the PAC learning model some of the algorithms, such as the Perceptron algorithm, did not initially appear to fall in the SQ framework but SQ analogues were later found for all known learning techniques except Gaussian elimination (for examples see [Kearns, 1998] and [Blum et al., 1997]). We expect the situation to be similar even in the broader context of search problems over distributions.

The STAT oracle we defined can return any value within the given tolerance and therefore can make adversarial choices. We also aim to prove lower bounds against algorithms that use a potentially more benign, "honest" statistical oracle. The honest statical oracle gives the algorithm the true value of a boolean query function on a randomly chosen sample. This model makes the sample complexity explicit and is based on the Honest SQ model in learning by Yang [2001] (which itself is based on an earlier model of Jackson [2003]).

Definition 2 (honest statistical algorithms)**.** *Let D be the input be a distribution over the domain X. An **honest statistical** algorithm does not have direct access to samples from D, but instead makes calls to an oracle* HSTAT_D*, which takes as input any function $h \in \mathcal{H} : X \to \{-1, 1\}$.* $\mathrm{HSTAT}_D(h)$ *takes an independent random sample x from D and returns $h(x)$.*

Note that the HSTAT oracle draws a fresh sample upon each time it is called. Without re-sampling each time, an honest statistical algorithm could easily recover the sample bit-by-bit, making it equivalent to the usual access to random samples. The **sample complexity** of an

honest statistical algorithm is defined to be the number of calls it makes to the HSTAT oracle. Note that the HSTAT oracle can be used to simulate STAT (with high probability) by taking the average of $O(1/\tau^2)$ replies of HSTAT for the same function[1] h. While it might seem that access to HSTAT gives an algorithm more power than access to STAT we will show that HSTAT can be simulated using STAT and also prove sample complexity lower bounds for honest statistical algorithms directly.

We are now ready to formally define problems over distributions.

Definition 3 (search problems over distributions)**.** *For a domain X, let $\mathcal{D}$ be a set of distributions over X, let $\mathcal{F}$ be a set of solutions and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map from a distribution $D \in \mathcal{D}$ to a subset of solutions $\mathcal{Z}(D) \subseteq \mathcal{F}$ that are defined to be valid solutions for D. The* search problem *$\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ is to find a valid solution $f \in \mathcal{Z}(D)$ given access to random samples from any $D \in \mathcal{D}$.*

We note that this definition captures decision problems by having $\mathcal{F} = \{0, 1\}$. With slight abuse of notation, for a solution $f \in \mathcal{F}$ we denote by $\mathcal{Z}^{-1}(f)$ the set of distributions in $\mathcal{D}$ for which f is a valid solution.

For some of the optimization problems we consider, it is natural to let the solution space $\mathcal{F}$ contain real-valued functions over X and define the valid functions $\mathcal{Z}(D) = \{f \in \mathcal{F} \mid \mathbf{E}_{x\sim D}[f(x)] \geq \mathbf{E}_{x\sim D}[f^*(x)] - \epsilon\}$, where $f^* \doteq \max_{f\in\mathcal{F}} \mathbf{E}_{x\sim D}[f(x)]$, i.e., the set of functions that are within additive error ϵ of being optimal. We refer to finding such a valid function as ϵ-**optimization**.

2.2 Statistical Dimension of search problems

The main tool in our analysis is an information-theoretic bound on the complexity of statistical algorithms based on the structure of a search problem over a distribution. Our definitions and techniques draw heavily upon the statistical query (SQ) model in learning theory, wherein the complexity of a large class of learning algorithms (most known learning algorithms) is characterized via a single parameter called the SQ dimension. Roughly speaking, it corresponds to the number of nearly uncorrelated labeling functions in the class [Blum et al., 1994, Kearns, 1998]. We introduce a natural generalization of this idea to search problems over arbitrary sets of distributions and prove a lower bound on the complexity of statistical algorithms based on the generalized notion. In addition, instead of relying on a bound on pairwise correlations, our dimension relies on a bound on average correlations in a large set of distributions. This weaker condition allows us to derive the tight bounds on the complexity of statistical algorithms for the planted k-clique problem.

We now define our dimension formally. For two functions $f, g : X \to \mathcal{R}$ and a distribution D with probability density function $D(x)$, the inner product of f and g over D is defined as

$$\langle f, g\rangle_D \doteq \mathop{\mathbf{E}}_{x\sim D}[f(x)g(x)].$$

The norm of f over D is $\|f\|_D = \sqrt{\langle f, f\rangle_D}$. We remark that, by convention, the integral from the inner product is taken only over the support of D, i.e. for $x \in X$ such that $D(x) \neq 0$. We also note that if $i = j$ above, the quantity $\langle \frac{D_i}{D} - 1, \frac{D_i}{D} - 1\rangle_D$ is known as the $\chi^2(D_i, D)$ distance. For a set $\mathcal{D}'$ of m distributions over X and a reference distribution D over X we define

$$\rho(\mathcal{D}', D) \doteq \frac{1}{m^2} \sum_{D_1, D_2 \in \mathcal{D}'} \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right|.$$

[1]Unlike HSTAT, STAT allows non-boolean functions that can be handled by first converting a real-valued query h to several boolean queries.

We are now ready to define the concept of statistical dimension.

Definition 4. *For $\bar{\gamma} > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, let d be the largest integer such that there exists a* reference *distribution D over X such that for every $f \in \mathcal{F}$ there exists a set of $m > 0$ distributions $\mathcal{D}_f = \{D_1, \ldots, D_m\} \subseteq \mathcal{D} \setminus \mathcal{Z}^{-1}(f)$ satisfying the following property: for any subset $\mathcal{D}' \subseteq \mathcal{D}_f$ where $|\mathcal{D}'| \geq m/d$, $\rho(\mathcal{D}', D) < \bar{\gamma}$. We define the* ***statistical dimension*** *with average correlation $\bar{\gamma}$ of $\mathcal{Z}$ to be d and denote it by* $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$.

The statistical dimension with average correlation $\bar{\gamma}$ of a search problem gives a lower bound on the complexity of any statistical algorithm for the problem that uses queries of tolerance $\sqrt{\bar{\gamma}}$.

Theorem 1. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\bar{\gamma} > 0$ let $d = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$. Any statistical algorithm requires at least d calls of tolerance $\tau = \sqrt{\bar{\gamma}}$ to the* STAT *oracle to solve $\mathcal{Z}$.*

It also gives a lower bound on the sample complexity of any honest statistical algorithm.

Theorem 2. *Let X be a domain and $\mathcal{Z}$ be a search problem over a class of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\bar{\gamma} > 0$ let $d = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$. Any honest statistical algorithm that solves $\mathcal{Z}$ with probability greater than* $13/14$ *requires at least*

$$\min\left\{\frac{1}{8\bar{\gamma}}, \frac{d}{100}\right\}$$

samples from HSTAT *oracle.*

The bound on the average correlation of large subsets upon which our notion is based can be easily obtained from a bound on pairwise correlations. Pairwise correlations are easier to analyze and therefore we now a define a special case of our statistical dimension based on pairwise correlations. This version can also be easily related to the statistical query dimension from learning theory (see Section 6). Hence, we define a second notion of statistical dimension, which is easier to work with in some cases.

Definition 5 (statistical dimension)**.** *For $\gamma, \beta > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. Let m be the maximum integer such that there exists a* reference *distribution D over X such that for every $f \in \mathcal{F}$ there exists a set of m distributions $\mathcal{D}_f = \{D_1, \ldots, D_m\} \subseteq \mathcal{D} \setminus \mathcal{Z}^{-1}(f)$ satisfying the following property:*

$$\left|\left\langle \frac{D_i}{D} - 1, \frac{D_j}{D} - 1 \right\rangle_D\right| \leq \begin{cases} \beta \text{ for } i = j \in [m] \\ \gamma \text{ for } i \neq j \in [m]. \end{cases}$$

We define the ***statistical dimension*** *with pairwise correlations (γ, β) of $\mathcal{Z}$ to be m and denote it by* $\mathrm{SD}(\mathcal{Z}, \gamma, \beta)$.

A corresponding lower bound can be obtained as a corollary of Theorem 1.

Corollary 1. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\gamma, \beta > 0$ let $m = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$. Any statistical algorithm requires at least $m(\tau^2 - \gamma)/(\beta - \gamma)$ calls of tolerance $\tau > 0$ to the* STAT *oracle to solve $\mathcal{Z}$.*

As we show in Section 3, this corollary follows by an appropriate choice of parameters. Furthermore, we can obtain a similar corollary for honest statistical algorithms (see Section 3.2, Corollary 4).

To conclude this section, we mention that in related work in the context of convex optimization, Raginsky and Rakhlin [2011] consider sequential optimization from noisy information and prove information-theoretic lower bounds.

2.3 Lower bounds

Our main lower bound is for the bipartite planted clique problem, for which we are able to show the following lower bound.

Theorem 3. *For any constant $\delta > 0$ and any $k \leq n^{1/2-\delta}$, at least $n^{\Omega(\log \log n)}$ queries of tolerance $\tau = \tilde{\Omega}(k/n)$ are required to find a planted bipartite clique of size k by any statistical algorithm.*

We note that this bound is close to tight. For every vertex in the clique, the probability that the corresponding bit of a randomly chosen point is set to 1 is $1/2+k/(2n)$ whereas for every vertex not in the clique this probability is $1/2$. Therefore using n queries of tolerance $k/(4n)$ it is easy to detect the planted clique.

We also give a sample complexity lower bound. To place this bound in context, we note that it is easy to detect whether a clique of size k has been planted using $O(n^2/k^2)$ samples: compute the average of $\sum_{i=1}^{n} x_i$; this will be noticeably higher if a clique has been planted. Moreover the clique subset itself can be found with this number of samples via the eigenvector approach. The next theorem is a lower bound that applies to any statistical algorithm. In particular, it implies that for cliques of size smaller than $\sqrt{n}$, one needs more than n samples for statistical algorithms to work.

Theorem 4. *For any constant $\delta > 0$ and any $k \leq n^{1/2-\delta}$, $\tilde{\Omega}(n^2/k^2)$ samples are required by any honest statistical algorithm to find a planted clique of size k.*

A closely related problem is the *planted densest subgraph* problem, where edges in the planted subset appear with higher probability than in the remaining graph. This is a variant of the densest k-subgraph problem, which itself is a natural generalization of k-clique that asks to recover the densest k-vertex subgraph of a given n-vertex graph [Feige, 2002, Khot, 2004, Bhaskara et al., 2010, 2012]. The conjectured hardness of its average case variant, the planted densest subgraph problem, has been used in public key encryption schemes [Applebaum et al., 2010] and in analyzing parameters specific to financial markets [Arora et al., 2010]. Our lower bounds extend in a straightforward manner to this problem.

We next turn to other applications of statistical dimension to some natural optimization problems over distributions. In particular, we show that any statistical algorithm for the moment maximization problem defined above, as well as distributional variants of MAX-XOR-SAT and k-CLIQUE must have high complexity.

Theorem 5. *For the rth moment maximization problem let $\mathcal{F}$ be the set of functions indexed by all possible unit vectors $u \in \mathcal{R}^n$, defined over the domain $\{-1,1\}^n$ with $f_u(x) = (u \cdot x)^r$. Let $\mathcal{D}$ be the set of all distributions over $\{-1,1\}^n$. Then for r odd and $\delta > 0$, at least $\tau^2(\binom{n}{r} - 1)$ queries of tolerance τ are required to $\left(\frac{r!}{2(r+1)^{r/2}} - \delta\right)$-optimize over $\mathcal{F}$ and $\mathcal{D}$ for any statistical algorithm.*

In words, any statistical algorithm that maximizes the r'th moment (for odd r) to within roughly $(r/e)^{r/2}$ must have complexity that grows as $\binom{n}{r}$.

The MAX-XOR-SAT problem over a distribution is defined as follows.

Problem 3 (MAX-XOR-SAT)**.** *Let D be a distribution over XOR clauses of arbitrary length, in n variables. The **MAX-XOR-SAT** problem is to find an assignment x that maximizes the number of satisfied clauses under the given distribution.*

In the worst case, it is known that MAX-XOR-SAT is NP-hard to approximate to within $1/2 - \epsilon$ for any constant ϵ [Håstad, 2001]. In practice, local search algorithms such as WalkSat [Selman et al., 1995] are commonly applied as heuristics for maximum satisfiability problems. We show that the distribution version of MAX-XOR-SAT is unconditionally hard for algorithms that locally seek to improve an assignment by flipping variables as to satisfy more clauses, giving some theoretical justification for the observations of Selman et al. [1995]. Moreover, our proof even applies to the case when there exists an assignment that satisfies all the clauses generated by the target distribution.

Theorem 6. *For the MAX-XOR-SAT problem, let $\mathcal{F}$ be the set of functions indexed by all possible assignments in n variables and whose domain is the set of all clauses. (The value that such a function takes when evaluated on a clause is the truth value of the clause under the given assignment.) Let $\mathcal{D}$ be the set of all distributions over clauses, then for $\delta > 0$, at least $\tau^2(2^n - 1)$ queries of tolerance τ are required to $\left(\frac{1}{2} - \delta\right)$-optimize over $\mathcal{F}$ and $\mathcal{D}$ for any statistical algorithm.*

Next, we consider the distribution version of the k-clique problem.

Problem 4 (distributional k-clique)**.** *Let D be a distribution over graphs G. The **k-clique** problem is to find a subset S of size k that maximizes the probability that S is a clique in G.*

Detecting whether a graph has a clique of size k is NP-Hard [Karp, 1972], fixed-parameter intractable (hard for W[1] [Downey and Fellows, 1999]) and no algorithm faster than $O(n^{.792k})$ is known [Nešetřil, 1985], even for a large constant k. While our lower bound does not give insight into the computational hardness of k-clique on worst-case inputs, it says that the k-clique problem over a distribution on graphs has high complexity for any statistical algorithm.

Theorem 7. *For the distributional k-clique problem, let $\mathcal{F}$ be the indicator functions indexed by subsets S of k vertices and whose domain is the set of all graphs on n vertices, that indicate whether S is a k-clique in the input graph. Let $\mathcal{D}$ be the set of distributions over graphs on n vertices. Then for $\delta > 0$, at least $\tau^2(\binom{n}{k} - 1)$ queries of tolerance τ are required to $\left(2^{-\binom{k}{2}} - \delta\right)$-optimize over $\mathcal{F}$ and $\mathcal{D}$ for any statistical algorithm.*

3 Lower bounds from Statistical Dimension

Here we prove the general lower bounds. In later sections, we will compute the parameters in these bounds for specific problems of interest.

3.1 Lower bounds for Statistical Algorithms

We begin with the proof of Theorem 1.

Proof of Theorem 1. Let $\mathcal{A}$ be a statistical algorithm that uses q queries of tolerance $\tau = \sqrt{\bar{\gamma}}$ to to solve $\mathcal{Z}$ over a class solutions $\mathcal{F}$ and class of distribution $\mathcal{D}$, such that $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}) = m$. Let D be the reference distribution for which the value d is achieved. We simulate $\mathcal{A}$ by answering any query $h : X \to [-1, 1]$ of $\mathcal{A}$ with value $\mathbf{E}_D[h(x)]$. Let $h_1, h_2, \ldots, h_q$ be the queries asked by $\mathcal{A}$ in this simulation and let f be the output of $\mathcal{A}$.

By definition of SDA, there exists a set of m distributions $\mathcal{D}_f = \{D_1, \ldots, D_m\}$ for which f is not a valid solution and such that for every $\mathcal{D}' \subseteq \mathcal{D}_f$, either $\rho(\mathcal{D}', D) < \bar{\gamma}$ or $|\mathcal{D}'| \leq m/d$.

In the rest of the proof for conciseness we drop the subscript D from inner products and norms.

To lower bound q, we use a generalization of an elegant argument of Szörényi [2009]. For every $k \leq q$ let A_k be the set of distributions D_i such that $|\mathbf{E}_D[h_k(x)] - \mathbf{E}_{D_i}[h_k(x)]| > \tau$. To prove the desired bound we first prove that following two claims:

1. $\sum_{k \leq q} |A_k| \geq m$;
2. for every k, $|A_k| \leq m/d$.

Combining these two immediately implies the desired bound $q \geq d$.

To prove the first claim we assume, for the sake of contradiction, that there exists $D_i \notin \cup_{k \leq q} A_k$. Then for every $k \leq q$, $|\mathbf{E}_D[h_k(x)] - \mathbf{E}_{D_i}[h_k(x)]| \leq \tau$. This implies that the replies of our simulation $\mathbf{E}_D[h_k(x)]$ are within τ of $\mathbf{E}_{D_i}[h_k(x)]$. By the definition of $\mathcal{A}$, this implies that f is a valid solution for $\mathcal{Z}$ on D_i, contradicting the condition that $D_i \in \mathcal{D} \setminus Z^{-1}(f)$.

To prove the second claim, suppose that $|A_k| > m/d$

$$\mathop{\mathbf{E}}_{D_i}[h_k(x)] - \mathop{\mathbf{E}}_{D}[h_k(x)] = \mathop{\mathbf{E}}_{D}\left[\frac{D_i}{D}\, h_k(x)\right] - \mathop{\mathbf{E}}_{D}[h_k(x)] = \left\langle h_k, \frac{D_i}{D} - 1 \right\rangle.$$

Let $\hat{D}_i(x) = \frac{D_i(x)}{D(x)} - 1$, (where the convention is that $\hat{D}_i(x) = 0$ if $D(x) = 0$). We will next show upper and lower bounds on the following quantity

$$\left\langle h_k, \sum_{i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i \rangle \right\rangle.$$

By Cauchy-Schwartz we have that

$$\begin{aligned}
\left\langle h_k, \sum_{i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i \rangle \right\rangle^2 &\leq \|h_k\|^2 \cdot \left\| \sum_{i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i \rangle \right\|^2 \qquad (1) \\
&\leq \|h_k\|^2 \cdot \left(\sum_{i,j \in A_k} \left| \langle \hat{D}_i, \hat{D}_j \rangle \right| \right) \\
&\leq \|h_k\|^2 \cdot \rho(A_k, D) \cdot |A_k|^2.
\end{aligned}$$

As before, we also have that

$$\begin{aligned}\left\langle h_k, \sum_{i\in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right\rangle^2 &= \left(\sum_{i\in A_k} \langle h_k, \hat{D}_i\rangle \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle\right)^2 \\ &= \left(\sum_{i\in A_k} (\mathsf{sign}\langle h_k, \hat{D}_i\rangle) \cdot \langle h_k, \hat{D}_i\rangle\right)^2 \\ &\geq \tau^2|A_k|^2 = \bar{\gamma}|A_k|^2. \end{aligned} \tag{2}$$

By combining these two inequalities we obtain that $\|h_k\|^2 \cdot \rho(A_k, D) \geq \tau^2$, which for $\|h_k\|^2 \leq 1$ implies that $\rho(A_k, D) \geq \bar{\gamma}$ which contradicts the definition of SDA. □

We now give the simple proof of the pairwise correlation version of statistical dimension-based lower bound (Corollary 1).

Proof of Corollary 1. Take $d = m(\tau^2 - \gamma)/(\beta - \gamma)$; we will prove that $\mathrm{SDA}(\mathcal{Z}, \tau^2) \geq d$ and apply Theorem 1. Consider a set of distributions $\mathcal{D}' \subset \mathcal{D}$ where $|\mathcal{D}'| \geq m/d = (\beta - \gamma)/(\tau^2 - \gamma)$:

$$\begin{aligned}\rho(\mathcal{D}', D) &= \frac{1}{|S|^2} \sum_{D_1, D_2 \in \mathcal{D}'} \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| \\ &\leq \frac{1}{|\mathcal{D}'|^2} \left(|S|\beta + (|\mathcal{D}'|^2 - |\mathcal{D}'|)\gamma\right) \\ &\leq \gamma + \frac{\beta - \gamma}{|\mathcal{D}'|} \\ &\leq \tau^2 \end{aligned}$$

□

We can also use the same way to bound the average correlation to obtain a direct bound on SDA using a bound on SD.

Corollary 2. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\gamma, \beta > 0$ let $m = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$. Then $\mathrm{SDA}(\mathcal{Z}, 2\gamma) \geq \frac{m\gamma}{\beta - \gamma}$.*

The next corollary shows a setting of the parameters that is useful for our applications in Section 5.

Corollary 3. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. If for $m > 0$, $\mathrm{SD}(\mathcal{Z}, \gamma = \frac{m^{-2/3}}{2}, \beta = 1) \geq m$ then at least $m^{1/3}/2$ calls of tolerance $m^{-1/3}$ to the* STAT *oracle are required to solve $\mathcal{Z}$.*

3.2 Lower bounds for Honest Statistical Algorithms

Next we address lower bounds for the HSTAT Oracle. The quantity $1/\tau^2$ can be thought of as representing the sample complexity of a statistical algorithm up to logarithmic factors. On one hand, q queries can be estimated to tolerance τ using $O(\log q/\tau^2)$ samples (with any constant probability of success). On the other hand, $\Omega(1/\tau^2)$ samples are necessary to estimate the expectation

of a "not-too-biased" query (with expectation bounded away from 1 and -1 by a constant) with constant probability of success. Strongly "biased" queries (such as a function which is identical to 1 on every sample) can be estimated with fewer samples to tolerance τ but our lower bound on the required tolerance can also be proportionately strengthened for such queries. We prove these points formally in our sample complexity lower bound for honest statistical algorithms. In this section it will be more convenient for us to assume that query functions used by the honest oracle are $\{0,1\}$ instead of $\{-1,1\}$. This does not change the model in any way since we can replace the value -1 with 0 in the query function and then replace 0 with -1 in the response. We will need the following two lemmas before proving Theorem 2.

Lemma 1. *For a query $h : X \to \{0,1\}$ and $\tau = \sqrt{\bar{\gamma}}$, let $A(h,\tau)$ be the set of distributions D_i in $\mathcal{D}_f$ such that $|\mathbf{E}_D[h(x)] - \mathbf{E}_{D_i}[h(x)]| > \tau\sqrt{\mathbf{E}_D[h(x)]}$,*

$$|A(h,\tau)| \leq m/d,$$

where D, $\bar{\gamma}$, $\mathcal{D}_f$, m and d are as defined in Theorem 1 and its proof.

Proof. In the proof of Theorem 1 we obtain that $|A_k| \leq m/d$ whenever $\|h_k\|^2 \cdot \rho(A_k, D) \geq \tau^2$. For $\tau = \sqrt{\bar{\gamma}}$, we can also obtain the same conclusion under the condition $\|h_k\|^2 \cdot \rho(A_k, D) \geq \tau^2 \cdot \|h_k\|^2$. In other words, we can obtain that $|A_k| \leq m/d$ also when A_k is defined as the set of distributions D_i such that $|\mathbf{E}_D[h_k(x)] - \mathbf{E}_{D_i}[h_k(x)]| > \tau \cdot \|h_k\|$. We now observe that for a $\{0,1\}$ function h_k, $||h_k||^2 = \mathbf{E}_D[h_k(x)]$. This implies that, in the notation of our lemma, $|A(h,\tau)| \leq m/d$. □

Lemma 2. *Let $X \sim B(1,p)$. Then, for any $p' \in (0,1)$,*

$$\mathop{\mathbf{E}}_{X}\left(\frac{\mathbf{Pr}[B(1,p) \text{ generated } X]}{\mathbf{Pr}[B(1,p') \text{ generated } X]}\right) = 1 + \frac{(p-p')^2}{p'(1-p')}.$$

Proof. If $X = 1$, the ratio is p/p' and when $X = 0$, then it is $(1-p)/(1-p')$. Thus, the expected ratio is

$$r = \frac{p^2}{p'} + \frac{(1-p)^2}{1-p'} = 1 + \frac{(p-p')^2}{p'(1-p')}.$$

□

We are now ready for the proof of the main lower bound.

Proof of Theorem 2. Our generative model for HSTAT's interaction with an algorithm is as follows: HSTAT picks as the target D with probability 1/2 and with probability 1/2 picks a D_i uniformly at random. Denote this random variable $\tilde{D}$. Upon a query of h_j, HSTAT draws a sample x_j from $\tilde{D}$, and responds with $h_j(x_j)$. After q rounds, the algorithm outputs its best guess of $\tilde{D}$. Because $\tilde{D}$ is drawn randomly, it makes sense to talk about the algorithm's success probability with respect to the randomness of $\tilde{D}$ and x_j.

An equivalent model is as follows: there is some joint distribution over $\tilde{D}$ and the possible responses of the HSTAT oracle. HSTAT will not choose $\tilde{D}$ first, but will answer queries according to their marginal distributions: when the algorithm presents query h_1, HSTAT returns an answer chosen according the marginal distribution of $h_1(x_1)$ (obtained by integrating out the $\tilde{D}$ variable). Subsequently, when the algorithm asks query h_j, HSTAT responds according to the marginal distribution of $h_j(x_j)$ conditioned on the previous responses $h_1(x_1), \ldots, h_{j-1}(x_{j-1})$. After the q^{th}

query HSTAT will pick $\tilde{D}$ from the marginal conditioned on $h_1(x_1), \ldots, h_q(x_q)$ and the algorithm will output a guess conditioned on $h_1(x_1), \ldots, h_q(x_q)$. It is clear that this is equivalent to the first model, but it captures the sources of randomness and available information much better. We call this the joint model, and will use it to prove our honest statistical algorithm lower bound.

Denote the result of the first j queries as $\omega_j = (h_1(x_1), \ldots, h_j(x_j))$, and let B denote an algorithm which outputs a guess based on ω_q: to maximise the probability that B's output and HSTAT's are the same:

$$\begin{aligned} \max_{B} \quad & \mathbf{Pr}[B(\omega_q) = \tilde{D} | \omega_q] \\ \text{s.t.} \quad & \sum_{D_i} \mathbf{Pr}[B(\omega_q) = D_i | \omega_q] = 1. \end{aligned}$$

We can rewrite the objective function as follows – B is adapted to ω_q and is independent of $\tilde{D}$.

$$\mathbf{Pr}[B(\omega_q) = \tilde{D} | \omega_q] = \sum_{D_i} \mathbf{Pr}[B(\omega_q) = D_i | \omega_q] \, \mathbf{Pr}[\tilde{D} = D_i | \omega_q].$$

The optimal B is deterministic and picks the D_i with greatest conditional probability. By construction, B has this quantity as its success probability. Since the algorithm can do no better than picking maximum conditional probabilities as its output, we will assume that it in fact does so. Clearly, making the algorithm more powerful still preserves any lower bounds.

We will analyze the conditional probability of D and show that this quantity never exceeds 7/8. The conditional probabilities can be rewritten by Bayes rule:

$$\mathbf{Pr}[D_i | h_1(x_1), \ldots, h_q(x_q)] = \frac{\mathbf{Pr}[h_1(x_1), \ldots, h_q(x_q) | D_i] \, \mathbf{Pr}[D_i]}{\mathbf{Pr}[h_1(x_1), \ldots, h_q(x_q)]}$$

Since the queries are adaptive, we define a random variable H_j for the choice of the j^{th} query. We can then expand the conditional probability term.

$$\mathbf{Pr}[h_1(x_1), \ldots, h_q(x_q) | D_i] = \prod_{j=1}^{q} \mathbf{Pr}[H_j = h_j | D_i, \omega_{j-1}, H_1, \ldots, H_{j-1}] \, \mathbf{Pr}[h_j(x_j) | D_i, \omega_{j-1}, H_1, \ldots, H_j]$$

The H_j random variables and $\mathbf{Pr}[h_1(x_1), \ldots, h_q(x_Q)]$ are the same for each D_i, so we suppress these as a constant c. The $h_j(x_j)$ are conditionally independent when D_i is fixed. In this case, each h_j is a Bernoulli random variable with bias p_j^i.

$$\mathbf{Pr}[h_j(x_j) | D_i] = (p_j^i)^{h_j(x_j)} (1 - p_j^i)^{1 - h_j(x_j)}$$

Therefore, the conditional probability is given by:

$$\mathbf{Pr}[D_i | h_1(x_1), \ldots, h_q(x_q)] = c \, \mathbf{Pr}[D_i] \prod_{j=1}^{q} (p_j^i)^{h_j(x_j)} (1 - p_j^i)^{1 - h_j(x_j)}$$

Let $\tau = \sqrt{\overline{\gamma}}$. Using Lemma 1, we can bound the size of $A(h_j, \tau)$ which consists of D_i's whose p_j^i are substantially different from that of D (which we shall denote by p_j). The number of D_i's in

the union of $A(h_j, \tau)$ is at most qm/d. Thus, with $q \leq d/100$, there are at least $99m/100$ such D_i's remaining.

For the remaining D_i's, we know that $|p_j^i - p_j| \leq \tau\sqrt{\mathbf{E}_D[h_j]} = \tau\sqrt{p_j}$. We can always assume that $p_j \leq 1/2$, since any query h such that $\mathbf{E}_D[h] > 1/2$ can be replaced with query $1 - h$ and the response then flipped by the algorithm. This implies that $|p_j^i - p_j| \leq \tau\sqrt{p_j} \leq \tau\sqrt{2p_j(1-p_j)}$. For every query j, we can now bound in expectation the increase in conditional probability using Lemma 2. The ratios change by at most

$$1 + \frac{(p_j^i - p_j)^2}{p_j(1-p_j)} \leq 1 + \frac{2\tau^2 p_j(1-p_j)}{p_j(1-p_j)} = 1 + 2\bar{\gamma}$$

in any round (in expectation). After q queries, the expected ratio is at most:

$$(1 + 2\bar{\gamma})^q \leq 1.5$$

for $q < 1/8\bar{\gamma}$. We can obtain concentration by using Markov's inequality. Hence, $q \geq 1/8\bar{\gamma}$. In particular, in relative terms, the conditional probability of D increases by a factor of at most 1.5. In particular, if we compare the conditional probability of D with the total conditional probability across all the other D_i, we obtain a comparison between $\mathbf{Pr}[D|h_1(x_1), \ldots, h_q(x_q)] \leq 3/4c$ and $\sum_{D_i \notin A} \mathbf{Pr}[D_i|h_1(x_1), \ldots, h_q(x_q)] \geq 99/200c$ which yields that the conditional probability of D is strictly less than $7/8$. Let A denote the algorithm's output, we have the following bounds

$$\begin{aligned}
\mathbf{Pr}[A = D \wedge \tilde{D} \neq D] + \mathbf{Pr}[A \neq D \wedge \tilde{D} \neq D] &= 1/2 \\
\mathbf{Pr}[A = D \wedge \tilde{D} = D] &\leq 1/2 \\
\mathbf{Pr}[A = D \wedge \tilde{D} = D] - 7\,\mathbf{Pr}[A = D \wedge \tilde{D} \neq D] &\leq 0.
\end{aligned}$$

By taking a linear combination of these constraints in the ratio $(1, 6/7, 1/7)$, we obtain the bound:

$$\mathbf{Pr}[A = D \wedge \tilde{D} = D] + \mathbf{Pr}[A \neq D \wedge \tilde{D} \neq D] \leq \frac{13}{14}$$

and that the success probability of the algorithm is bounded by $13/14$.

Thus,

$$q \geq \min\left(\frac{1}{8\bar{\gamma}}, \frac{d}{100}\right)$$

□

We conclude this section with an application of Corollary 2 to obtain a version of Theorem 2 for the simpler (pairwise) version of statistical dimension.

Corollary 4. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\gamma, \beta > 0$ let $m = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$. Any deterministic honest statistical algorithm requires at least*

$$\min\left(\frac{1}{16\gamma}, \frac{1}{60}\sqrt{\frac{m}{\beta}}\right)$$

samples from HSTAT *oracle to solve $\mathcal{Z}$.*

Proof. We have

$$\max_{\gamma_0 \geq \gamma}\left(\min\left\{\frac{1}{16\gamma_0}, \frac{1}{100}\frac{m\gamma_0}{\beta - \gamma_0}\right\}\right)$$

Solving for γ_0 and substituting, we get our bound. □

3.3 Reductions between STAT and HSTAT

We now show that access to the honest statistical oracle is essentially equivalent to access to STAT. It has been observed in the context of learning [Yang, 2001] that, given a boolean query function h one can obtain an estimate of $\mathbf{E}_D[h]$ using $t = O(\log(1/\delta)/\tau^2)$ honest samples which with probability at least $1-\delta$ will be within τ of $\mathbf{E}_D[h]$. We also allow real-valued query functions in our model but any such query function can be replaced by $\lceil \log(1/\tau) \rceil + 2$ boolean queries each or tolerance $\tau/2$. A query i computes bit i of $1 + h(x) \in [0,2]$ so only $\lceil \log(1/\tau) \rceil + 2$ bits are necessary to get the value of $h(x)$ within $\tau/2$. Combining these two observations gives us the following theorem.

Theorem 8. *Let $\mathcal{Z}$ be a search problem and let A be a statistical algorithm that solved $\mathcal{Z}$ using q queries of tolerance τ. For any $\delta > 0$, there exists an honest statistical algorithm A' that uses at most $O(q \log(q/(\delta\tau))/\tau^2)$ samples and solves $\mathcal{Z}$ with probability at least $1-\delta$.*

We also show a reduction in the other direction, namely that the STAT oracle can be used to simulate the HSTAT oracle.

Theorem 9. *Let $\mathcal{Z}$ be a search problem and let A be an honest statistical algorithm that solved $\mathcal{Z}$ with probability at least δ using q samples from* HSTAT*. For any δ' there exists a statistical algorithm A' that uses at most q queries of tolerance $2 \cdot \delta'/q$ and solves $\mathcal{Z}$ with probability at least $\delta(1-\delta')$.*

Proof. A' simulates A as follows. Let $h_1 : X \to \{-1, 1\}$ be the first query of A and let $p = \mathbf{E}_{x \sim D}[h(x)]$. By asking the query $\mathrm{STAT}_D(h_1, \tau)$, for $\tau = \delta'/q$ we can get a value $p' \in [p-\tau, p+\tau]$. We flip a ± 1 coin with bias p' (that is one that outputs 1 with probability $(1+p')/2$ and -1 with probability $(1-p')/2$. We return the outcome to A. One can think of the coin flip with bias p' as the coin flip with bias p and then a correction with probability $|p'-p|/2$. Namely, if $p' > p$ then -1 is output with probability $(p'-p)/2$ and otherwise 1 is output with probability $(p-p')/2$. This implies that our simulation can be seen as an honest simulation with a random correction step that happens with probability at most $|p-p'|/2 \leq \tau/2 = \delta'/q$.

We continue the simulation of the rest of A' queries analogously. By the union bound, the probability of a correction step happening during the simulation (and hence of our simulation differing from the honest one) is at most δ', independently of other random events. Therefore A' is successful with probability at least $\delta(1-\delta')$. □

4 Planted Clique

We now prove the lower bound claimed in Theorem 3 on the problem of determining whether the given distribution on vectors from $\{0,1\}^n$ is just uniform or from a planted k-clique distribution as defined above.

For a subset $S \subseteq [n]$, let D_S be the distribution with a planted clique on the subset S. Let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k. For $i \in [m]$ we use D_i to denote D_{S_i}. The reference distribution in our lower bounds will be the uniform distribution over $\{0,1\}^n$ and let $\hat{D}_S$ denote $D_S/D - 1$. In order to apply our lower bounds based statistical dimension with average correlation we now prove that for the planted clique problem average correlations of large sets must be small. We start with a lemma that bounds the correlation of two planted clique distributions relative to the reference distribution D as a function of the overlap between the cliques.

Lemma 3. *For $i, j \in [m]$,*

$$\left\langle \hat{D}_i, \hat{D}_j \right\rangle_D \leq \frac{2^\lambda k^2}{n^2},$$

where $\lambda = |S_i \cap S_j|$.

Proof. For the distribution D_i, we consider the probability $D_i(x)$ of generating the vector x. Then,

$$D_i(x) = \begin{cases} (\frac{n-k}{n})\frac{1}{2^n} + (\frac{k}{n})\frac{1}{2^{n-k}} & \text{if } \forall \lambda \in S, x_\lambda = 1 \\ (\frac{n-k}{n})\frac{1}{2^n} & \text{otherwise.} \end{cases}$$

Now we compute the vector $\hat{D}_i = \frac{D_i}{D} - 1$:

$$\frac{D_i}{D} - 1 = \begin{cases} \frac{k2^k}{n} - \frac{k}{n} & \text{if } \forall \lambda \in S, x_\lambda = 1 \\ -\frac{k}{n} & \text{otherwise.} \end{cases}$$

We then bound $\left\langle \hat{D}_i, \hat{D}_j \right\rangle_D$

$$\begin{aligned} \left\langle \hat{D}_i, \hat{D}_j \right\rangle_D &\leq \frac{2^{n-2k+\lambda}}{2^n}\left(\frac{k2^k}{n} - \frac{k}{n}\right)^2 + 2\left(\frac{2^{n-k}}{2^n}\right)\left(\frac{k2^k}{n} - \frac{k}{n}\right)\left(-\frac{k}{n}\right) + \left(-\frac{k}{n}\right)^2 \\ &\leq \frac{2^\lambda k^2}{n^2} \end{aligned}$$

□

We now give a bound on the average correlation of any $\hat{D}_S$ with a large number of distinct clique distributions.

Lemma 4. *For $\kappa < 1/2$ and $k \leq n^\kappa$, let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k and $\{D_1, \ldots, D_m\}$ be the corresponding distributions on $\{0,1\}^n$. Then for any integer $\ell \leq k$, set S of size k and subset $A \subseteq \{S_1, \ldots, S_m\}$ where $|A| \geq 4(m-1)/n^{\ell(1-2\kappa)}$,*

$$\frac{1}{|A|}\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle < 2^{\ell+2}\frac{k^2}{n^2}.$$

Proof. In this proof we first show that if the total number of sets in A is large then most of sets in A have a small overlap with S. We then use the bound on the overlap of most sets to obtain a bound on the average correlation of D_S with distributions for sets in A.

Formally, we let $\alpha = \frac{k^2}{n^2}$ and using Lemma 3 get the bound $\langle \hat{D}_i, \hat{D}_j \rangle \leq 2^{|S_i \cap S_j|}\alpha$. Summing over $S_i \in A$,

$$\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle \leq \sum_{S_i \in A} 2^{|S \cap S_i|}\alpha.$$

For any set $A \subseteq \{S_1, \ldots, S_m\}$ of size t this bound is maximized when the sets of A include S, then all sets that intersect S in $k-1$ indices, then all sets that intersect S in $k-2$ indices and so on until the size bound t is exhausted. We can therefore assume without loss of generality that A is defined in precisely this way.

Let

$$T_\lambda = \{S_i \mid |S \cap S_i| = \lambda\}$$

denote the subset of all k-subsets that intersect with S in exactly λ indices. Let λ_0 be the smallest λ for which $A \cap T_\lambda$ is non-empty. We first observe that for any $1 \le j \le k-1$,

$$\frac{|T_j|}{|T_{j+1}|} = \frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{k}{j+1}\binom{n-k}{k-j-1}} = \frac{(j+1)(n-2k+j+1)}{(k-j-1)(k-j)} \ge \frac{(j+1)(n-2k)}{k(k+1)} \ge \frac{(j+1)n^{1-2\kappa}}{2}.$$

By applying this equation inductively we obtain,

$$|T_j| \le \frac{2^j \cdot |T_0|}{j! \cdot n^{(1-2\kappa)j}} < \frac{2^j \cdot (m-1)}{j! \cdot n^{(1-2\kappa)j}}$$

and

$$\sum_{k \ge \lambda \ge j} |T_\lambda| < \sum_{k \ge \lambda \ge j} \frac{2^\lambda \cdot (m-1)}{\lambda! \cdot n^{(1-2\kappa)\lambda}} \le \frac{4(m-1)}{n^{(1-2\kappa)j}} \ .$$

By definition of λ_0, $|A| \le \sum_{j \ge \lambda_0} |T_j| < 4(m-1)/n^{(1-2\kappa)\lambda_0}$. In particular, if $|A| \ge 4(m-1)/n^{\ell(1-2\kappa)}$ then $n^{\ell(1-2\kappa)}/4 < n^{(1-2\kappa)\lambda_0}/4$ or $\lambda_0 < \ell$. Now we can conclude that

$$\begin{aligned}
\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle &\le \sum_{j=\lambda_0}^{k} 2^j |T_j \cap A| \alpha \\
&\le \left(2^{\lambda_0} |T_{\lambda_0} \cap A| + \sum_{j=\lambda_0+1}^{k} 2^j |T_j| \right) \alpha \\
&\le \left(2^{\lambda_0} |T_{\lambda_0} \cap A| + 2 \cdot 2^{\lambda_0+1} |T_{\lambda_0+1}| \right) \alpha \\
&\le 2^{\lambda_0+2} |A| \alpha < 2^{\ell+2} |A| \alpha.
\end{aligned}$$

To derive the last inequality we need to note that for every $j \ge 0$, $2^j|T_j| > 2(2^{j+1}|T_{j+1}|)$ we can therefore telescope the sum. □

Lemma 4 gives a simple way to bound the statistical dimension with average correlation of the planted bipartite k-clique problem.

Theorem 10. *For $\kappa < 1/2$ and $k \le n^\kappa$ let $\mathcal{Z}$ the planted bipartite k-clique problem. Then for any $\ell \le k$,* $\mathrm{SDA}(\mathcal{Z}, 2^{\ell+2}k^2/n^2) \ge n^{\ell(1-2\kappa)}/4$.

Proof. Let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k and $\mathcal{D} = \{D_1, \ldots, D_m\}$ be the corresponding distributions on $\{0,1\}^n$. For every solution $S \in \mathcal{F}$, $\mathcal{Z}^{-1}(S) = D_S$ and let $\mathcal{D}_S = \mathcal{D} \setminus \{D_S\}$. Note that $|\mathcal{D}_S| = m - 1$.

Let $\mathcal{D}'$ be a set of distributions $\mathcal{D}' \subseteq \mathcal{D}_S$ such that $|\mathcal{D}'| \ge 4(m-1)/n^{\ell(1-2\kappa)}$. Then by Lemma 4, for every $S_i \in \mathcal{D}'$,

$$\frac{1}{|\mathcal{D}'|} \sum_{S_j \in \mathcal{D}'} \langle \hat{D}_i, \hat{D}_j \rangle < 2^{\ell+2} \frac{k^2}{n^2}.$$

In particular, $\rho(\mathcal{D}', D) < 2^{\ell+2}\frac{k^2}{n^2}$. By the definition of SDA, this means that $\mathrm{SDA}(\mathcal{Z}, 2^{\ell+2}k^2/n^2) \ge n^{\ell(1-2\kappa)}/4$. □

Theorems 1 and 10 imply the following corollary, as well as Theorem 3.

Corollary 5. *For any $\kappa < 1/2$, $k \leq n^\kappa$ and any $\ell \leq k$ at least $n^{\ell(1-2\kappa)}/4$ queries of tolerance $\tau = 2^{\ell/2+1}\frac{k}{n}$ are required to solve the planted bipartite k-clique problem. In particular, for any constant κ and $\ell = \log\log n$ we obtain that $n^{\Omega(\log\log n)}$ queries of tolerance $\tau = \tilde{\Omega}(k/n)$ are required.*

Theorems 2 and 10 also imply the sample complexity lower bound stated in Theorem 4.

4.1 Planted Densest Subgraph

We will now show the lower bound on detecting a planted densest subset, a generalization of the planted clique problem.

Problem 5 (planted bipartite densest k-subgraph)**.** *For $1 \leq k \leq n$, let $S \subseteq \{1, 2, \ldots, n\}$ be a set of k vertex indices and D_S be a distribution over $\{0,1\}^n$ such that when $x \sim D_S$, with probability $1-(k/n)$ the entries of x are chosen uniformly and independently from $\{0,1\}$, and with probability k/n the k coordinates in S are each, independently, set to 1 with probability $q > 1/2$ and the rest are chosen uniformly and independently from $\{0,1\}$. The **planted bipartite densest k-subgraph** problem is to find the unknown subset S given access to samples from D_S.*

We note that when $p = 1$ this is equivalent to the planted clique problem. For this problem, we are able to prove the following bound.

Lemma 5. *Let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k for $k \leq n^\kappa$ for $\kappa < 1/2$ and $\ell \leq k$ with associated planted densest subgraph distributions $\{D_1, \ldots, D_m\}$. Then for any set S of size k and subset $A \subseteq \{S_1, \ldots, S_m\}$ where $|A| \geq m/d$,*

$$\frac{1}{|A|}\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle \leq 8\left(\left(2(q^2(1-q)^2)\right)^{\frac{\log(2m/d)}{(1-2\kappa)\log(n)}} - 1\right)\frac{k^2}{n^2}.$$

Proof. Our planted sets on n coordinates will be of size k, with pairwise overlap λ as before. The difference is to consider the probability q (as opposed to 1) of edges appearing in the plant in the calculation. We define

$$\xi_S(x, q) \doteq q^{|S\cap x|}(1-q)^{k-|S\cap x|}$$

and consider

$$D_S(x) = \frac{k}{n}\left(\frac{\xi_S(x,q)}{2^{n-k}}\right) + \frac{n-k}{n}\left(\frac{1}{2^n}\right).$$

Then the quantity

$$\frac{D_S(x)}{D} - 1 = 2^k\frac{k}{n}(\xi_S(x,q)) - \frac{k}{n}.$$

We need to compute

$$\begin{aligned}
\left\langle \frac{D_S(x)}{D} - 1, \frac{D_{S_i}(x)}{D} - 1 \right\rangle_D &= \frac{1}{2^n}\left(\frac{k}{n}\right)^2 \sum_{x\in\{0,1\}^n} \left(2^k\xi_S(x,q) - 1\right)\left(2^k\xi_S(x,q) - 1\right) \\
&= \frac{1}{2^n}\left(\frac{k}{n}\right)^2 \sum_{x\in\{0,1\}^n} \left(2^{2k}\xi_S(x,q)\xi_{S_i}(x,q) - 2\cdot 2^k\xi_S(x,q) + 1\right) \\
&= \frac{1}{2^n}\left(\frac{k}{n}\right)^2 \left(2^{n+\lambda} \sum_{x\in\{0,1\}^{2k-\lambda}} \xi_S(x,q)\xi_{S_i}(x,q) \right. \\
&\qquad \left. - 2\cdot 2^k 2^{n-k} \sum_{x\in\{0,1\}^k} \xi_S(x,q) + 2^n\right) \\
&= \frac{1}{2^n}\left(\frac{k}{n}\right)^2 \left(2^{n+\lambda}(q^2+(1-q)^2)^\lambda - 2^n\right) \\
&= \frac{k^2}{n^2}2^\lambda(q^2+(1-q)^2)^\lambda - \frac{k^2}{n^2}
\end{aligned}$$

The rest of the proof proceeds as in Lemma 4, except that with the same choice of λ_0, we obtain

$$\begin{aligned}
\sum_{S_i\in A} \langle \hat{D}_S, \hat{D}_i\rangle &\le \sum_{j=\lambda_0}^{k} \alpha(2^j(q^2+(1-q)^2)^j - 1)|T_j\cap A| \\
&\le \left(|T_{\lambda_0}\cap A|(2^{\lambda_0}(q^2+(1-q)^2)^j - 1) + \sum_{j=\lambda_0+1}^{k} 2^j|T_j|(q^2+(1-q)^2)^j - \sum_{j=\lambda_0+1}^{k}|T_j|\right)\alpha \\
&\le \left(2^{\lambda_0}|T_{\lambda_0}\cap A|((q^2+(1-q)^2)^j - 1)\right. \\
&\qquad \left. + 2\cdot 2^{\lambda_0+1}|T_{\lambda_0+1}|(q^2+(1-q)^2)^{\lambda_0+1} - \sum_{j=\lambda_0+1}^{k}|T_j|\right)\alpha \\
&\le 8(2^{\lambda_0}(q^2+(1-q)^2)^{\lambda_0} - 1)|A|\alpha.
\end{aligned}$$

□

Theorem 11. *For $\kappa < 1/2$ and $k \le n^\kappa$ let $\mathcal{Z}$ the planted bipartite densest subgraph problem. Then for any $\ell \le k$, $q > 1/2$,*

$$\mathrm{SDA}\left(\mathcal{Z}, 8\left((2(q^2(1-q)^2))^{\frac{\log(2m/d)}{(1-2\kappa)\log(n)}} - 1\right)\right) \ge n^{\ell(1-2\kappa)}/4.$$

With appropriate choices of parameter settings, we get the following Corollary.

Corollary 6. *For constants $c, \delta > 0$, density $q \le 1/2 + 1/n^c$, and $k \le n^{1/2-\delta}$, any honest statistical algorithm requires $\tilde{\Omega}((n^{2+c})/k^2)$ samples to find a planted densest subgraph of size k.*

5 Other applications of Statistical Dimension

In this section, we use Definition 5 together with the bound in Corollary 1 to get unconditional lower bounds for a variety of optimization problems. A recurring concept in our constructions will be a **parity function**, χ. We first explore some properties of parity functions.

Definition 6 (parity)**.** *For $x \in \{0,1\}^n$ and $c \in \{0,1\}^n$, let $\chi_c : \{0,1\}^n \to \{-1,1\}$.*

$$\chi_c(x) \doteq -(-1)^{c \cdot x}.$$

Namely, $\chi_c(x) = 1$ if $c \cdot x$ is odd, and -1 otherwise.

Note: *for convenience*[2]*, we will sometimes use $x \in \{\pm 1\}^n$, in which case we abuse notation and define $\chi_c(x) = -\prod_{i:\ c_i=1} x_i$. This corresponds to the embedding of x from $\{0,1\} \to \{-1,1\}$ of $0 \to 1$, $1 \to -1$.*

Further, we define distributions uniform over the examples classified positive by a parity.

Definition 7 (distributions D_c)**.** *Let $x \in \{\pm 1\}^n$ and $c \in \{0,1\}^n$ and let $S_c = \{x \mid \chi_c(x) = 1\}$. We define D_c to be the uniform distribution over S_c.*

Lemma 6. *For $c \in \{0,1\}^n$, $c \neq \bar{0}$ and the uniform distribution U over $\{-1,1\}^n$, the following hold:*

$$1) \quad \mathop{\mathbf{E}}_{x \sim D_c}[\chi_{c'}(x)] = \begin{cases} 1 \text{ if } c = c' \\ 0 \text{ otherwise.} \end{cases} \qquad 2) \quad \mathop{\mathbf{E}}_{x \sim U}[\chi_c(x)\chi_{c'}(x)] = \begin{cases} 1 \text{ if } c = c' \\ 0 \text{ otherwise.} \end{cases}$$

Proof. To show Part 1) note that if $c = c'$ then $\mathbf{E}_{x \in S_c}[\chi_c(x)] = 1$. If $c \neq c' \neq \bar{0}$ then it is easy to see that $|S_c \cap S_{c'}| = |S_c|/2 = |S_{c'}|/2$ and so $\mathbf{E}_{x \in S_c}[\chi_{c'}(x)] = \sum_{x \in S_c \cap S_{c'}} 1 + \sum_{x \in S_c \setminus S_{c'}} (-1) = 0$. Part 2) states the well-known fact that the parity functions are uncorrelated relative to the uniform distribution. □

These two facts will imply that when $D = U$ (the uniform distribution) and the D_i's consist of the D_c's, we can set $\gamma = 0$ and $\beta = 1$, when considering the statistical dimension of the problems presented in the following sections.

5.1 MAX-XOR-SAT

We first formalize the MAX-XOR-SAT problem introduced in Problem 3. Let D be a distribution over XOR clauses $c \in \{0,1\}^n$. We interpret $c_i = 1$ as variable i appearing in c and otherwise not; for simplicity, no variables are negated in the clauses. The problem is to find an assignment $x \in \{0,1\}^n$ that maximizes the expected number of satisfied XOR clauses. We now give the statistical dimension of this problem, from which Theorem 6 follows.

Theorem 12. *For the MAX-XOR-SAT problem, let $\mathcal{F} = \{\chi_x\}_{x \in \{0,1\}^n}$, let $\mathcal{D}$ be the set of all distributions over clauses $c \in \{0,1\}^n$, and for any $\delta > 0$, let $\mathcal{Z}$ be the problem of $(\frac{1}{2} - \delta)$-optimizing over $\mathcal{F}$ and $\mathcal{D}$. Then $\mathrm{SD}(\mathcal{Z}, 0, 1) \geq 2^n - 1$.*

[2]For the moment maximization problem, it is necessary for our argument that examples x be $\in \{-1,1\}^n$, whereas for MAX-XOR-SAT, the argument is much cleaner when x is in $\{0,1\}^n$. It is, therefore, natural to use the same notation for the corresponding parity problems.

Proof. Maximizing the expected number of satisfied clauses is equivalent to maximizing the quantity

$$\max_{x\in\{0,1\}^n} \mathop{\mathbf{E}}_{c\sim D}[\chi_x(c)].$$

This proof is a fairly direct application of Lemma 6 to the definition of statistical dimension.

For the conditions in Definition 5, for each each of the 2^n possible assignments to x let D_x be the uniform distribution over the clauses $c \in \{0,1\}^n$ such that $\chi_c(x) = 1$.

Because $\chi_c(x)$ is symmetric in x and c, the conditions in Definition 5, with $\beta = 1$ and $\gamma = 0$, which follow from Lemma 6, are satisfied for the 2^n distributions D_c, with $D = U$. Because $\chi_c(x) = 1$ when assignment x satisfies clause c and -1 otherwise, we need to scale the approximation term by $1/2$ when measuring the *fraction* of satisfied clauses □

Corollary 7. *Any statistical algorithm for a MAX-XOR-SAT instance asymptotically requires $2^{n/3}$ queries of tolerance $2^{-n/3}$ to find an assignment that approximates the maximum probability of satisfying clause drawn from an unknown distribution to less than an additive term of $1/2$.*

5.2 k-Clique

We first formalize the distributional k-clique problem. Let D be a distribution over $X = \{0,1\}^{\binom{n}{2}}$, corresponding to graphs G on n vertices. For $G \in X$, let

$$I_S(G) \doteq \begin{cases} 1 & \text{if } S \text{ induces a clique in } G \\ 0 & \text{otherwise.} \end{cases}$$

The k-clique problem is to find a subset $S \subseteq V$ of size k that maximizes $\mathbf{E}_{G\sim D}[I_S(G)]$.

We now give the statistical dimension of distributional k-clique, from which Theorem 7 follows.

Theorem 13. *For the distributional k-clique problem, let $\mathcal{F} = \{I_S\}_{|S|=k}$, let $\mathcal{D}$ be the set of distributions over graphs on n vertices, and for any $\delta > 0$, let $\mathcal{Z}$ be the problem of $\left(2^{-\binom{k}{2}} - \delta\right)$-optimizing over $\mathcal{F}$ and $\mathcal{D}$. Then $\mathrm{SD}(\mathcal{Z}, 0, 1) \geq \binom{n}{k} - 1$.*

Proof. We shall compute the statistical dimension of distributional k-clique with $\epsilon = 2^{-\binom{k}{2}} - \delta$ (for $\delta > 0$), $\gamma = 0$, and $\beta = 1$ and show it is $\binom{n}{k}$.

For any subset of edges $T \in V \times V$, and graph $G \in X$, we can define the function

$$\mathrm{parity}_T(G,k) \doteq \begin{cases} 1 \text{ if } |E(G) \cap T| \text{ has the same parity as } \binom{k}{2} \\ -1 \text{ otherwise.} \end{cases}$$

Note that $\mathrm{parity}_T(G,k) = (-1)^{\left(|E(G)\cap T| \, + \binom{k}{2}\right)}$.

As both T and G lie in $\{0,1\}^{\binom{n}{2}}$, note that $\mathrm{parity}_T(G,k)$ is simply $\chi_T(G)$ or (its negation, depending on k). Let $T_1, \ldots, T_d$ be all the $\binom{n}{k}$ cliques on k vertices. We generate the distributions $D_1, \ldots, D_d$ so that D_i is uniform on the graphs G such that $|E(G) \cap T_i| = \binom{k}{2} \mod 2$. The distribution D is the uniform over all graphs G. By Lemma 6, these choices justify $\beta = 1, \gamma = 0$.

We notice that the set of vertices of the clique T_i maximizes $\mathbf{E}_{D_i}[I_S(G)]$ while the set of edges of the clique maximizes $\mathbf{E}_{D_i}[\mathrm{parity}_T(G,k)]$, namely we have that

$$V(T_i) = \arg\max_{S\in V:|S|=k} \left(\mathop{\mathbf{E}}_{G\sim D_i}[I_S(G)] \right) = V\left(\arg\max_{T\in V\times V} \left(\mathop{\mathbf{E}}_{G\sim D_i}[\mathrm{parity}_T(G,k)] \right) \right).$$

By definition $\mathbf{E}_{G\sim D_i}[\text{parity}_T(G,k)] \leq 1$, with equality iff $T = T_i$.

For $S_i = V(T_i)$ we have that $I_{S_i}(G) = 1$ iff T_i is a clique in G. Since any setting of the edges not in T_i appears equiprobably under D_i and since there are $2^{\binom{k}{2}-1}$ possible settings for edges between vertices in $V(T_i)$ occurring equiprobably in graphs from D_i, it follows that $\mathbf{E}_{G\sim D_i}[I_{S_i}(G)] = 2^{-\binom{k}{2}+1}$.

On the other hand, if $S_j \neq V(T_i)$ then all subsets of edges among the vertices of S_j appear equiprobably under D_i. Hence, for $j \neq i$, $\mathbf{E}_{G\sim D_i}[I_{S_j}] = 2^{-\binom{k}{2}}$, as only 1 of every $2^{\binom{k}{2}}$ subgraphs on k vertices forms a clique. This allows us to set $\epsilon = 2^{-\binom{k}{2}} - \delta$, for any $\delta > 0$.

Because our distributions were generated by the k vertex subsets, we have shown the statistical dimension to be $\binom{n}{k} - 1$. □

Corollary 8. *Any statistical algorithm for a k-clique instance asymptotically requires $\binom{n}{k}^{1/3}$ queries of tolerance $\binom{n}{k}^{-1/3}$ to find an assignment that approximates the maximum probability of satisfying clause drawn from an unknown distribution to less than an additive term of $2^{-\binom{k}{2}}$.*

5.3 Moment Maximization

We recall the moment maximization problem. Let D be a distribution over $\{-1,1\}^n$ and let $r \in Z^+$. The moment maximization problem is to find a unit vector u that maximizes $\mathbf{E}_{x\sim D}[(u\cdot x)^r]$.

Before going to the main proof, we need to prove a property of odd moments.

Lemma 7. *Let $r \in Z^+$ be odd and let $c \in \{0,1\}^n$. Let D_c be the distribution uniform over $x \in \{-1,1\}^n$ for which $\chi_c(x) = -1$. Then, $\forall u \in \mathcal{R}^n$,*

$$\mathop{\mathbf{E}}_{x\sim D_c}[(x\cdot u)^r] = r! \prod_{i:\ c_i=1} u_i.$$

Proof. From Lemma 8 we have that

$$\forall u \in \mathcal{R}^n, \mathop{\mathbf{E}}_{x\sim D_c}[(x\cdot u)^r] = r! \prod_{i:\ c_i=1} u_i + \mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[(x\cdot u)^r]. \tag{3}$$

the lemma follows now since when r is odd

$$\mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[(x\cdot u)^r] = \mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[((-x)\cdot u)^r] = 0.$$

□

Lemma 8. *Under the conditions of Lemma 7,*

$$\forall u \in \mathcal{R}^n, \mathop{\mathbf{E}}_{x\sim D_c}[(x\cdot u)^r] = r! \prod_{i:\ u_i=1} u_i + \mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[(x\cdot u)^r]. \tag{4}$$

Proof. Notice that

$$\mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[(x\cdot u)^r] = \frac{1}{2}\mathop{\mathbf{E}}_{\chi_c(x)=-1}(x\cdot u)^r + \frac{1}{2}\mathop{\mathbf{E}}_{\chi_c(x)=1}(x\cdot u)^r$$

and that

$$-\mathop{\mathbf{E}}_{x\in\{\pm1\}^n}\chi_c(x)(x\cdot u)^r=\frac{1}{2}\mathop{\mathbf{E}}_{\chi_c(x)=-1}(x\cdot u)^r-\frac{1}{2}\mathop{\mathbf{E}}_{\chi_c(x)=1}(x\cdot u)^r,$$

therefore

$$\mathop{\mathbf{E}}_{x\sim D_c}[(x\cdot u)^r]=\mathop{\mathbf{E}}_{x\in\{\pm1\}^n}[(x\cdot u)^r]-\mathop{\mathbf{E}}_{x\in\{\pm1\}^n}\chi_c(x)[(x\cdot u)^r].$$

Equation 4 follows now by Lemma 9 below. □

Lemma 9. *Let c be an r parity on the variables indexed by set $I=\{i_1,\ldots,i_r\}$, $c\in\{0,1\}^n$. Let u be an arbitrary vector in $\mathcal{R}^n$. Then*

1. $\mathbf{E}_{x\in\{\pm1\}^n}=\mathbf{E}[\chi_c(x)(x\cdot u)^i]=0$ *for* $i<r$
2. $\mathbf{E}_{x\in\{\pm1\}^n}=\mathbf{E}[\chi_c(x)(x\cdot u)^r]=-r!\prod_{i\in I}u_i$.

Proof. To prove Part 1, we have

$$\begin{aligned}\mathbf{E}\left[\chi_c(x)(x\cdot c)^i\right] &= \mathbf{E}\left[\chi_c(x)\sum_{t_1+\ldots t_n=i}\binom{i}{t_1,\ldots,t_n}\prod_{i\in[r]}(u_ix_i)^{t_i}\right]\\ &= \sum_{t_1+\ldots t_n=i}\binom{i}{t_1,\ldots,t_n}\mathbf{E}\left[\chi_c(x)\prod_{i\in[r]}(u_ix_i)^{t_i}\right].\end{aligned}$$

Notice that if there is some variable $j\in I$ such that $t_j=0$ then $\mathbf{E}_x[\chi_c(x)\prod_{i\in[r]}(u_ix_i)^{t_i}]=0$, as the term corresponding to x always cancels out with the term corresponding to the element obtained by flipping the jth bit of x. Since $i<r$ every term $\prod_{i\in[r]}(u_ix_i)^{t_i}$ must contain some $t_j=0$ with $j\in I$, which concludes that $\mathbf{E}[\chi_c(x)(x\cdot c)^i]=0$.

To prove Part 2 of the lemma, we will induct on n. For $n=r$,

$$\begin{aligned}\mathbf{E}\left[\chi_c(x)(x\cdot u)^r\right] &= \mathbf{E}\left[\chi_c(x)\sum_{t_1+\ldots t_r=r}\binom{r}{t_1,\ldots,t_r}\prod_{i\in[r]}(u_ix_i)^{t_i}\right]\\ &= \sum_{t_1+\ldots t_r=r}\binom{r}{t_1,\ldots,t_r}\mathbf{E}\left[\chi_c(x)\prod_{i\in[r]}(u_ix_i)^{t_i}\right].\end{aligned}$$

As before, if some $t_j=0$ and $j\in I=[r]$ then $\mathbf{E}[\chi_c(x)\prod_{i\in[r]}(u_ix_i)^{t_i}]=0$, since for each x and $\tilde{x}$ obtained by flipping the jth bit of x it is the case that $\chi_c(x)=-\chi_c(\tilde{x})$. Therefore

$$\begin{aligned}\mathbf{E}\left[\chi_c(x)(x\cdot u)^r\right] &= \mathbf{E}\left[\chi_c(x)\binom{r}{1,1,\ldots,1}\prod u_ix_i\right]\\ &= -r!(\prod u_i)\,\mathbf{E}\left[\prod x_i^2\right]\\ &= -r!(\prod u_i).\end{aligned}$$

Assume now the identity holds for n. Let $c\in\{0,1\}^{n+1}$ and let $j\notin I$, and for $x\in\{0,1\}^{n+1}$ define $x_{-j}\in\{0,1\}^n$ to be x with the jth bit punctured.

Then

$$\begin{aligned}
\mathbf{E}\left[\chi_c(x)(x\cdot u)^r\right] &= \mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j}+x_ju_j)^r\right] \\
&= \mathbf{E}\left[\chi_c(x)\sum_{0\le i\le r}\binom{r}{i}(x_{-j}\cdot u_{-j})^{r-i}(x_ju_j)^i\right] \\
&= \mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^r\right]+\mathbf{E}\left[\chi_c(x)\sum_{1\le i\le r}\binom{r}{i}(x_{-j}\cdot u_{-j})^{r-i}(x_ju_j)^i\right] \\
&= -r!\prod_{i\in I}u_i+\sum_{1\le i\le r}\binom{r}{i}\mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}(x_ju_j)^i\right]. \qquad (5)
\end{aligned}$$

If i is even then

$$\mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}(x_ju_j)^i\right]=(u_j)^i\,\mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}\right]=0$$

by Part 1 of the lemma. If i is odd then

$$\begin{aligned}
\mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}(x_ju_j)^i\right] &= u_j^i\,\mathbf{E}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}x_j\right] \\
&= u_j^i\frac{1}{2}\left(\mathop{\mathbf{E}}_{x_j=1}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}\right]-\mathop{\mathbf{E}}_{x_j=-1}\left[\chi_c(x)(x_{-j}\cdot u_{-j})^{r-i}\right]\right) \\
&= 0,
\end{aligned}$$

since $j\notin I$ and so $\chi_c(x)=\chi_c(\tilde{x})$, where $\tilde{x}$ is obtained from x by flipping the jth bit. We can now conclude that Equation (5) $=-r!\prod_{i\in I}u_i$. □

Corollary 9. *Let $r\in Z^+$ be odd*[3] *and let $c\in\{0,1\}^n$. Let D_c be the distribution uniform over $x\in\{-1,1\}^n$ for which $\chi_c(x)=-1$. Then, $\mathbf{E}_{x\sim D_c}[(x\cdot u)^r]$ is maximized when $u=r^{-1/2}c$.*

Proof. From Lemma 7, clearly whenever $c_i=0$, we have $u_i=0$. It follows from the AM-GM inequality that the product is maximized when the remaining coordinates are equal. □

Now we are ready to show the statistical dimension of moment maximization, from which Theorem 5 follows.

Theorem 14. *For the rth moment maximization problem let $\mathcal{F}=\{(u\cdot x)^r\}_{u\in\mathcal{R}^n}$ and let $\mathcal{D}$ be a set of distributions over $\{-1,1\}^n$. Then for an odd r and $\delta>0$, let $\mathcal{Z}$ denote the problem of $\left(\frac{r!}{2(r+1)^{r/2}}-\delta\right)$-optimizing over $\mathcal{F}$ and $\mathcal{D}$. Then $\mathrm{SD}(\mathcal{Z},0,1)\ge\binom{n}{r}-1$.*

Proof. Let $D_1,\ldots,D_d$ be distributions where D_i is uniform over all examples x in $\{0,1\}^n$, where such that $\chi_{c_i}(x)=1$; this again allows us to consider $\beta=1$ and $\gamma=0$.

Corollary 9 shows that under the distribution D_i, the moment function

$$\max_{u\in\mathcal{R}:\|u\|=1}\mathop{\mathbf{E}}_{x\sim D_i}\left[(u\cdot x)^r\right]$$

[3]This statement does not hold for r even.

is maximized at $u = r^{-1/2}c$. So, to maximize the moment, one equivalently needs to find the correct target parity.

To compute the needed ϵ, for r odd, Lemma 7 tells us that the expected moment is simply $r! \prod_{i:c_i=1} u_i$, and for unit vectors, is maximized when $\forall i : c_i = 1,\ u_i = r^{-1/2}$ (and $u_i = 0$ for the other coordinates). This yields a maximum moment of $(r!)r^{-r/2}$ for any D_i.

In comparison, if the measured moment is equal to $(r!)(r+1)^{-r/2}$, a simple consequence of Lemma 7 is that to minimize $\sum_{i:c_i=1} u_i^2$, then for all i s.t. $c_i = 1$, we have $u_i = (r+1)^{-1/2}$. Hence, for all i s.t. $c_i = 0$, u_i cannot take value greater than $1 - r((r+1)^{-1/2})^2 = (r+1)^{-1/2}$, implying a moment of at most $(r!)(r+1)^{-r/2}$ on $D_{c'}$. This gives a bound of

$$\epsilon \ \geq\ (r!)r^{-r/2} - (r!)(r+1)^{-r/2} \ \geq\ \frac{r!}{2(r+1)^{r/2}}.$$

The $\binom{n}{r}$ parities generating the different distributions give the statistical dimension. □

Corollary 10. *For r odd, any statistical algorithm for the moment maximization problem asymptotically requires $\binom{n}{r}^{1/3}$ queries of tolerance $\binom{n}{r}^{-1/3}$ to approximate the r-th moment to less than an additive term of $\frac{r!}{2(r+1)^{r/2}}$.*

6 Relationship to Statistical Queries in learning

We will now use Corollary 1 to demonstrate that our work generalizes the notion of statistical query dimension and the statistical query lower bounds from learning theory. In an instance of a PAC learning problem, the learner has access to random examples of an unknown boolean function $f' : X' \to \{-1, 1\}$ from a set of boolean functions $\mathcal{C}$ (whenever necessary, we use $'$ to distinguish variables from the identically named ones in the context of general search problems). A random example is a pair including a point and its label $(x', c(x'))$ such that x' is drawn randomly from an unknown distribution D'. For $\epsilon > 0$, the goal of an ϵ-accurate learning algorithm is to find, with high probability, a boolean hypothesis h' for which $\mathbf{Pr}_{x' \sim D'}[h'(x') \neq f'(x')] \leq \epsilon$.

A statistical query (SQ) learning algorithm [Kearns, 1998] has access to a statistical query oracle for the unknown function f' and distribution D' in place of random examples. A query to the SQ oracle is a function $\phi : X' \times \{-1, 1\} \to [-1, 1]$ that depends on both the example x' and its label ℓ. To such a query the oracle returns a value v which is within τ of $\mathbf{E}_{D'}[\phi(x', c(x')]$, where τ is the tolerance parameter. A SQ algorithm does not depend on the randomness of examples and hence must always succeed.

Blum et al. [1994] defined the *statistical query dimension* or SQ-DIM of a set of functions $\mathcal{C}$ and distribution D' over X' as follows (we present a simplification and strengthening due to Yang [2005]).

Definition 8 (Blum et al. [1994])**.** *For a concept class $\mathcal{C}$ and distribution D', SQ-DIM$(\mathcal{C}, D') = d'$ if d' is the largest value for which there exist d' functions $c_1, c_2, \ldots, c_{d'} \in \mathcal{C}$ such that for every $i \neq j$, $|\langle c_i, c_j \rangle_{D'}| \leq 1/d'$.*

Blum et al. [1994] proved that if a class of functions is learnable using only a polynomial number of statistical queries of inverse polynomial tolerance then its statistical query dimension is polynomial. Yang [2005] strengthened their result and proved the following bound (see [Szörényi, 2009] for a simpler proof).

Theorem 15 (Yang [2005])**.** *Let $\mathcal{C}$ be a class of functions and D' be a distribution over X' and let $d' = SQ\text{-}DIM(\mathcal{C}, D')$. Then any SQ learning algorithm for $\mathcal{C}$ over D' that makes q queries of tolerance $1/d'^{1/3}$ and outputs an ϵ-accurate hypothesis for $\epsilon \leq 1/2 - 1/(2d'^{1/3})$ satisfies that $q \geq d'^{1/3}/2 - 1$.*

In this result the distribution D' is fixed and known to the learner (such learning is referred to as *distribution-specific*) and it can be used to lower bound the complexity of learning $\mathcal{C}$ even in a weak sense. Specifically, when the learning algorithm is only required to output a hypothesis h' such that $\mathbf{Pr}_{x' \sim D'}[h'(x') \neq c(x')] \leq 1/2 + \gamma'$ for some inverse polynomial γ' (or $\epsilon \leq 1/2 - \gamma'$).

We now claim that we can cast this learning problem as an optimization problem, and by doing so, we will obtain that our statistical dimension implies a lower bound on learning which is stronger than that of Yang [2005].

Let $\mathcal{L} = (\mathcal{C}, D', \epsilon)$ be an instance of a distribution-specific learning problem of a class of functions $\mathcal{C}$ over distribution D' to accuracy $1 - \epsilon$. We define the following 2ϵ-optimization problem $\mathcal{Z}_{\mathcal{L}}$ over distributions. The domain is all the labeled points or $X = X' \times \{-1, 1\}$. When the target function equals $c \in C$ the learning algorithm gets samples from the distribution D_c over X, where $D_c(x', c(x')) = D'(x')$ and $D_c(x', -c(x')) = 0$. Therefore we define the set of distributions over which we optimize to be $\mathcal{D}_{\mathcal{L}} = \{D_c \mid c \in \mathcal{C}\}$. Note that STAT oracle for D_c with tolerance τ is equivalent to the statistical query oracle for c over D' with tolerance τ. We can take the class of functions $\mathcal{F}_{\mathcal{L}}$ over which a learning algorithm optimizes to be the set of all boolean functions over X of the form $f(x', \ell) = f'(x') \cdot \ell$ for some boolean function f' over X' (an efficient learning algorithm can only output circuits of polynomial size but this distinction is not important for our information-theoretic bounds). We define $\mathcal{Z}_{\mathcal{L}}$ to be the problem of 2ϵ-optimizing over $\mathcal{F}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{L}}$. Note that for $f \in \mathcal{F}_{\mathcal{L}}$ and $D_c \in \mathcal{D}_{\mathcal{L}}$,

$$\mathop{\mathbf{E}}_{D_c}[f(x)] = \mathop{\mathbf{E}}_{D'}[f'(x') \cdot c(x')] = 1 - 2 \mathop{\mathbf{E}}_{D'}[f'(x') \neq c(x')]$$

and therefore learning to accuracy $1 - \epsilon$ is equivalent to 2ϵ-optimizing over $\mathcal{F}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{L}}$.

We claim that SQ-DIM$(\mathcal{C}, D')$-based lower bound given in Theorem 15 is effectively just a minor learning-specific simplification of our statistical dimension lower bound for $\mathcal{Z}_{\mathcal{L}}$ (Cor. 1).

Theorem 16. *Let $\mathcal{C}$ be a class of functions and D' be a distribution over X' and let $d' = SQ\text{-}DIM(\mathcal{C}, D')$. Denote by $\mathcal{L} = (\mathcal{C}, D', \epsilon)$ the instance of learning $\mathcal{C}$ over D' for $\epsilon = 1/2 - 1/(2d'^{1/3})$. Then*

$$\mathrm{SD}(\mathcal{Z}_{\mathcal{L}}, \gamma = 1/d', \beta = 1) \geq \left(d' - \frac{1}{1/d'^{2/3} - 1/d'}\right).$$

Proof. Let $c_1, c_2, \ldots, c_{d'}$ be the almost uncorrelated functions in $\mathcal{C}$ implied by the definition of SQ-DIM$(\mathcal{C}, D')$. We define the reference distribution D as the distribution for which for every $(x', \ell) \in X$, $D(x', \ell) = D'(x')/2$. We note that this ensures that $D(x', \ell)$ is non-vanishing only when $D'(x')$ is non-vanishing and hence the function $\left(\frac{D_c}{D} - 1\right)$ will be well-defined for all $c \in \mathcal{C}$. For every $c \in \mathcal{C}$, we have

$$\frac{D_c(x', c(x'))}{D(x', c(x'))} - 1 = 2 - 1 = 1 \quad \text{and} \quad \frac{D_c(x', -c(x'))}{D(x', -c(x'))} - 1 = 0 - 1 = -1.$$

Therefore, $\frac{D_c(x',\ell)}{D(x',\ell)} = \ell \cdot c(x')$. This implies that for any $c_i, c_j \in \mathcal{C}$,

$$\left\langle \frac{D_{c_i}}{D} - 1, \frac{D_{c_j}}{D} - 1 \right\rangle_D = \mathop{\mathbf{E}}_{D}[\ell \cdot c_i(x') \cdot \ell \cdot c_j(x')] = \mathop{\mathbf{E}}_{D'}[c_i(x') \cdot c_j(x')] = \langle c_i, c_j \rangle_{D'}.$$

Hence

1. for any $c \in \mathcal{C}$, $\left\| \frac{D_c(x)}{D(x)} - 1 \right\|_D^2 = 1$;
2. for any $i \neq j \leq d'$, $\left\langle \frac{D_{c_i}(x)}{D(x)} - 1, \frac{D_{c_j}(x)}{D(x)} - 1 \right\rangle_D \leq 1/d'$.

These properties imply that d' functions in $\mathcal{C}$ give d' distributions in $\mathcal{D}_{\mathcal{L}}$ whose distinguishing functions are almost uncorrelated. This is essentially the condition required to obtain a lower bound of d' on $\mathrm{SD}(\mathcal{Z}_{\mathcal{L}}, 1/d', 1)$. The only issue is that we need to exclude distributions for which any given $f \in \mathcal{F}_{\mathcal{L}}$ is 2ϵ-optimal. We claim that it is easy to bound the number of distribution which are 2ϵ-optimal for a fixed $f(x', \ell) = f'(x') \cdot \ell$ and whose distinguishing functions are almost uncorrelated. First, note that the condition of 2ϵ-optimality of f for D_c states that

$$\mathop{\mathbf{E}}_{D_c}[f(x)] \geq 1 - 2\epsilon \geq 1/d'^{1/3}.$$

On the other hand, $\mathbf{E}_D[f(x)] = 0$ and therefore $\mathbf{E}_{D_c}[f(x)] - \mathbf{E}_D[f(x)] \geq 1/d'^{1/3}$. This implies that if we view $f(x)$ as a query function then expectations of the query function relative to D_c and D differ by at least $\tau = 1/d'^{1/3}$. In the proof of Corollary 1, we proved that this is possible for at most $(\beta - \gamma)/(\tau^2 - \gamma)$ distributions with pairwise correlations (γ, β). For our parameters this gives a bound of $\frac{1}{1/d'^{2/3} - 1/d'}$ distributions. Hence for $m = d' - \frac{1}{1/d'^{2/3} - 1/d'}$ we obtain that for every $f \in \mathcal{F}_{\mathcal{L}}$ there exist m distributions $D_1, \ldots, D_m \subseteq \{D_{c_1}, \ldots, D_{c_{d'}}\} \setminus \mathcal{Z}_{\mathcal{L}}^{-1}(f)$ such that

1. for any $i \leq m$, $\left\| \frac{D_i(x)}{D(x)} - 1 \right\|_D = 1$;
2. for any $i \neq j \leq m$, $\left\langle \frac{D_i(x)}{D(x)} - 1, \frac{D_j(x)}{D(x)} - 1 \right\rangle_D \leq 1/d'$.

□

Applying Corollary 1, we get the following lower bound, which is twice larger than the $d'^{1/3}/2 - 1$ bound of Yang [2005].

Corollary 11. *Let $\mathcal{C}$ be a class of functions and D' be a distribution over X', let $d' = SQ\text{-}DIM(\mathcal{C}, D')$ and let $\epsilon = 1/2 - 1/(2d'^{1/3})$. Then any SQ learning algorithm requires at least $d'^{1/3} - 2$ queries of tolerance $1/d'^{1/3}$ to ϵ-accurately learn $\mathcal{C}$ over D'.*

6.1 Honest Statistical Queries

We now turn to the Honest SQ model [Jackson, 2003, Yang, 2001], which inspired our notion of statistical sampling algorithms. In the Honest SQ model, the learner has access to an HSQ oracle and can again evaluate queries which are a function of the data points and their labels. As in our HSTAT oracle, the queries are evaluated on an "honest" sample drawn from the target distribution. More precisely, the HSQ oracle accepts a function $\phi : X' \times \{-1, 1\} \to \{-1, 1\}$ and a sample size

$t > 0$, draws $x'_1, \ldots, x'_t \sim D'$, and returns the value $\frac{1}{t}\sum_{i=1}^{t} \phi(x', c(x'))$. The total sample count of an algorithm is the sum of the sample sizes it passes to HSQ.

We note that using our one-sample-per-query-function oracle HSTAT one can simulate estimation of queries from larger samples in the straightforward way while obtaining the same sample complexity. Therefore HSQ is equivalent to our HSTAT oracle.

We first observe that our direct simulation in Theorem 9 implies that the Honest SQ learning model is equivalent (up to polynomial factors) to the SQ learning model. We are not aware of this observation having been made before (although Valiant [2009] implicitly uses it to show that evolvable concept classes are also learnable in the SQ model).

We now show that using Corollary 4 we can derive sample complexity bounds on honest statistical query algorithms for learning.

Corollary 12. *Let $\mathcal{C}$ be a class of functions, D' be a distribution over X', $d' = SQ\text{-}DIM(\mathcal{C}, D')$ and $\epsilon = 1/2 - 1/(2d'^{1/3})$. Then the sample complexity of any Honest SQ algorithm for ϵ-accurate learning of $\mathcal{C}$ over D' is $\tilde{\Omega}(\sqrt{d'})$.*

This recovers the bound in Yang [2005] up to polynomial factors.

Acknowledgments

We thank Avrim Blum, Ravi Kannan, Michael Kearns, and Avi Wigderson for helpful discussions.

References

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *SODA*, pages 594–598, 1998.

Benny Applebaum, Boaz Barak, and Avi Wigderson. Public-key cryptography from different assumptions. In *STOC*, pages 171–180, 2010.

Sanjeev Arora, Boaz Barak, Markus Brunnermeier, and Rong Ge. Computational complexity and information asymmetry in financial products (extended abstract). In *ICS*, pages 49–65, 2010.

Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: an $o(n^{1/4})$ approximation for densest k-subgraph. In *STOC*, pages 201–210, 2010.

Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou. Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph. In *SODA*, pages 388–405, 2012.

A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *STOC*, pages 253–262, 1994.

Charles S. Brubaker. Extensions of principal component analysis. *Phd. Thesis, School of CS, Georgia Tech*, 2009.

S. Brubaker and Santosh Vempala. Random tensors and planted cliques. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 5687, pages 406–419. 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer-Verilag, 1999.

John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Math. Program.*, 114(1):101–114, 2008.

Uriel Feige. Relations between average case complexity and approximation complexity. In *IEEE Conference on Computational Complexity*, page 5, 2002.

Alan M. Frieze and Ravi Kannan. A new approach to the planted clique problem. In *FSTTCS*, pages 187–198, 2008.

A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48:798–859, July 2001. ISSN 0004-5411.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, 40(1):79–91, 2011.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np hard. *CoRR*, abs/0911.1393, 2009.

J. Jackson. On the efficiency of noise-tolerant PAC algorithms derived from statistical queries. *Annals of Mathematics and Artificial Intelligence*, 39(3):291–313, November 2003.

Mark Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 3(4):347–360, 1992.

Ari Juels and Marcus Peinado. Hiding cliques for cryptographic security. *Des. Codes Cryptography*, 20(3):269–280, 2000.

Ravi Kannan. personal communication.

Richard Karp. Reducibility among combinatorial problems. In R. Miller and J. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

Subhash Khot. Ruling out ptas for graph min-bisection, densest subgraph and bipartite clique. In *FOCS*, pages 136–145, 2004.

Scott Kirkpatrick, D. Gelatt Jr., and Mario P. Vecchi. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

Ludek Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2-3):193–212, 1995.

Frank McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Lorenz Minder and Dan Vilenchik. Small clique detection and approximate nash equilibria. 5687: 673–685, 2009.

Jaroslav; P. Nešetřil. On the complexity of the subgraph problem. *Commentationes Mathematicae Universitatis Carolinae*, 26(2):415–419, 1985.

M. Raginsky and A. Rakhlin. Information-Based Complexity, Feedback and Dynamics in Convex Programming. *Information Theory, IEEE Transactions on*, 57(10):7036–7056, October 2011. ISSN 0018-9448.

Bart Selman, Henry Kautz, and Bram Cohen. Local search strategies for satisfiability testing. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 521–532, 1995.

Balázs Szörényi. Characterizing statistical query learning: Simplified notions and proofs. In *ALT*, pages 186–200, 2009.

M Tanner and W Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Leslie G. Valiant. Evolvability. *J. ACM*, 56(1), 2009.

V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, January 1985. ISSN 0022-3239.

Ke Yang. On learning correlated boolean functions using statistical queries. In *Proceedings of ALT*, pages 59–76, 2001.

Ke Yang. New lower bounds for statistical query learning. *J. Comput. Syst. Sci.*, 70(4):485–509, 2005.